BIOLOGICAL CRITERIA
Technical Guidance for Survey Design and Statistical Evaluation of Biosurvey Data

CHAPTER 3 Designing the Sample Survey

The design of the sample survey is a critical element in the environmental assessment process, and certain statistical methods are associated with specific designs. This chapter examines various types of survey design and shows how the selection of the design relates to the interpretation and use of data within the biocriteria program. For information on designs not covered in this chapter, see Cochran, 1963; Cochran and Cox, 1957; Green, 1979; Williams, 1978; and Reckhow and Stow, 1990.

Efforts to design sample surveys frequently result in situations that force the investigator to evaluate the trade-offs between an increase in uncertainty and the costs of reducing this uncertainty (Reckhow and Chapra, 1983). But major components of uncertainty, including variability, error, and bias in biological and statistical sources, can sometimes be controlled by a well-specified survey design.

For example, variability can be caused by natural fluctuations in biological indicators over space and time; error can be associated with inaccurate data acquisition or reduction; and bias can occur when the sample is not representative of the population under review or when the samples are not randomly collected. These sources of uncertainty should be evaluated before the sampling design is selected because the best design will minimize the effects of variability, error, and bias on decision making.

Critical Aspects of Survey Design

Data collection within the biocriteria program requires the investigator to address issues associated with both *classical* and *experimental* survey designs. In general, experimental survey design focuses on the collection of data that leads to the testing of a specific hypothesis. Classical survey design is motivated less by hypothesis testing than by the "survey" concept. That is, the investigator gathers a relatively small amount of data, the sample, and extrapolates from it a view of the totality of available information.

In this chapter, we will address issues that overlap these design types. In addition, we will focus on designs appropriate to local, site-specific situations. For larger geographic survey designs, see Hunsaker and Carpenter (1990), or Linthurst et al. (1986).

Variability

A critical aspect of sampling design is to identify and separate components of variability, including the important ones of time, space, and random errors. Yearly and seasonal variations and spatial variations like those caused by changes in geographic patterns should be accounted for in the survey design. A design that stratifies the sampling based on knowledge of spatial and temporal changes in the abundance and character of biological indicators is preferred to systematic random sampling. That is, if biological indicators are known to exhibit temporal and spatial patterns, then sampling locations and times must be adjusted to match the biological variability.

Representativeness and Sampling Techniques

The object of a biological survey design is to reduce the total information available to a small sample: observations are made and data collected on a relatively small number of biological variables. Representativeness is, therefore, a key consideration in the design of sample collection procedures. The data generated during the survey should be representative of the population or process under evaluation. Biased samples occur when the data are not representative of the population. For example, a sample mean may be low (biased) because the investigator failed to sample geographic areas of high abundance.

Several techniques can increase the odds of collecting a representative sample; however, the technique most frequently used is *random sampling*. Theoretically, in simple random sampling, every unit in the population has the same chance of being included in the sample. Random sampling is a physical way to introduce independence among environmental measurements. In addition, random sampling has the affect of minimizing various types of bias in the interpretation of results.

If the geographic area sampled is large, with known or suspected environmental patterns, a good technique is to divide the area into relatively homogenous sections and randomly sample within each one. This technique is known as *stratified sampling*. Samples can be allocated to each section in proportion to the size of the area or to the known abundance of organisms within each area. In still other cases, *systematic sampling* may be appropriate. Systematic sampling improves precision in the sample estimates, especially when known spatial patterns exist

(Cochran, 1963). Randomly allocated replicate samples collected on a grid allow for good spatial coverage of patchy environments, yet minimize the potential for sampling bias.

Cause and Effect

In classical statistical experiments, a population is identified and randomly divided into two groups. The treatment is administered to one group; the other group serves as the control. The difference in the average response between the two groups indicates the effect of the treatment, and the random assignment of individuals to the groups permits an inference of causality because the observed difference results from the treatment and not from some preexisting difference between the groups.

In an ecological assessment, the treatment and control groups are not selected at random from a larger population, since the impacted site cannot be selected at random. And no matter how carefully the reference site is matched, the investigator cannot compensate for the lack of random selection. In this sense, a statistically valid test of the hypothesis that an observed difference between an impacted site and a control site results from a specific cause is impossible. The hypothesis that the two sites are different can be tested, but the difference cannot be attributed to a specific cause. In statistical terms, the stress on the impacted site is completely confounded with preexisting differences between the impact and reference site.

Although a firm case can be made that a site is subject to adverse impacts, investigators must realize that the site is an experimental unit that cannot be replicated. They must take care to avoid "pseudoreplication" (Hurlbert, 1984) — the testing of a hypothesis about adverse effects without appropriate statistical design or analysis methods. The problem is a misunderstanding or misspecification of the hypothesis being tested. It is avoided by understanding that only the hypothesis of a difference between sites can be statistically tested. Cause-and-effect issues cannot be resolved using statistical methods. Of course, establishing that a difference exists is an essential step in the process of demonstrating an adverse ecological effect. If there is no detectable difference, there is no cause to establish.

Methods used to establish causality can make use of statistical techniques, such as regression or correlation. For example, regression can be used to show that toxicity increases along with the concentration of some chemical known to originate from a wastewater outfall. The regression describes the relationship; it does not imply the cause, though presence of a strong relationship is evidence that a link exists.

One way to resolve these issues is to collect both spatial and temporal data from a control site. If the spatial control is missing and only before and after impact samples are available at the impacted site, statistical tests cannot rule out the possibility that the change would have occurred with or without the impact. If the temporal control is missing, the statistical tests cannot rule out the possibility that the differences between the control and impact site may have occurred prior to the impact. In practice, control data are rarely available in both spatial and temporal dimensions. Therefore, most environmental assessments detect only that differences exist between the control and impact sites. The causal link is more difficult to discern.

Controls

In environmental assessments, control or reference data are used in hypothesis tests to evaluate whether data from the control and impact site are statistically different. Evidence of impact is based on changes in the biological community that did not occur in the control area. Sources of control information include baseline data, reference site data, and numeric standards. The case for causality can be strengthened if the controls are properly selected.

In an ideal study design, both temporal and spatial control data should be collected (Green, 1979). The control site should be geographically separated from the impacted site but have similar physical and ecological features (e.g., elevation, temperature, wind patterns, and habitat type and disturbance). In aquatic habitats, parameters such as stream order, flow rate, and stream hydrography should be considered. Ideally, biological indicators of impact should be collected at the control site before and after the impact occurs.

Statistically, a valid control site should have conservative properties. That is, its statistics should be the same as at the impacted site except for the effects of the impact. Physical, chemical, and ecological variables should be measured and statistically evaluated to confirm that the impact and control sites are properly matched. Investigators should test for mean differences as well as differences in distribution. In addition, the variance of the physical and ecological similarities between the control and impact sites should be the same over time. For example, if the mean pH between the two sites is consistent but the impact site experiences much wider swings in pH than the control site, then the ability to confidently detect an impact for a pH-dependent toxicant is compromised. Samples within the control and reference site should be randomly allocated at some level. For example, in a random sampling design (Fig. 3.1), the

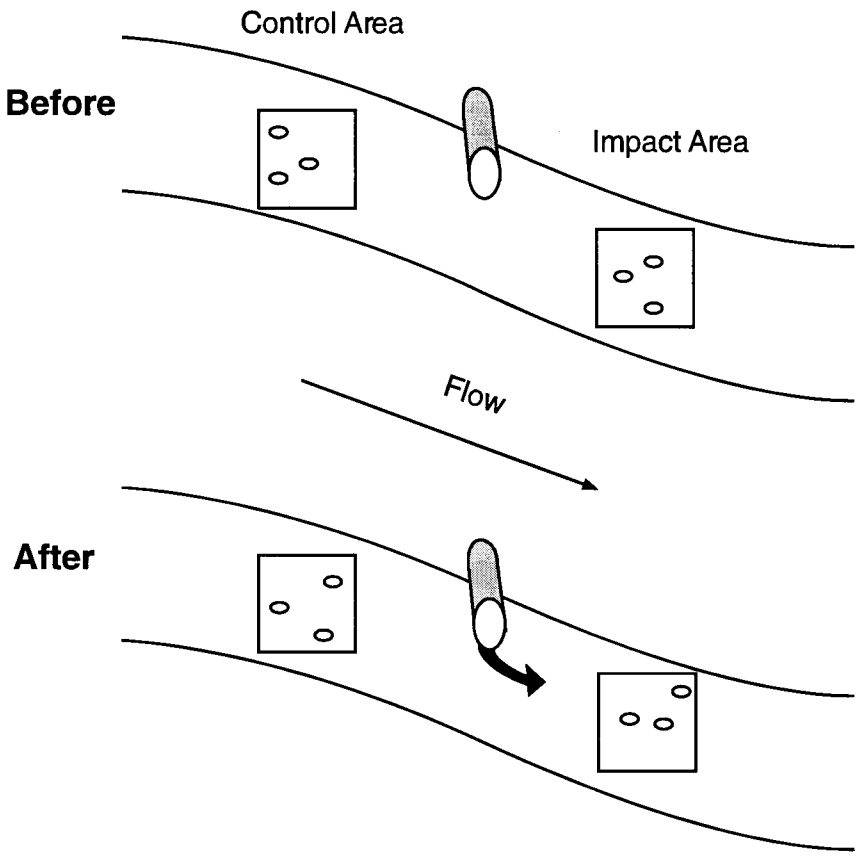

Figure 3.1—Random before and after control impact (BACI) sample design having both temporal and spatial dimensions. Random samples indicated are from within areas identified as being of similar habitat. (Adapted from Green, 1979.)

samples would be randomly allocated in a temporal/spatial framework that would allow for a number of different statistical analyses, including analysis of variance (ANOVA).

In an optimal study design, the impact would be in the future. Thus, baseline data providing a temporal control would be available to the investigator. In practice, baseline data are rarely available, and the investigator cannot be certain whether differences between the impact and control sites preceded or followed the impact. Therefore, cause and effect cannot be determined. However, the fact that a difference exists allows the investigator to hypothesize a causal link.

In some cases, biological variables collected at an impact site may be compared to a fixed numeric value rather than to a set of identical measurements collected at a reference site. Nevertheless, the issues associated with demonstrating causality remain the same. In addition, the investigator should note that the numeric criterion has no variance. It is usually presented as a single number with no associated uncertainty. In such cases, a *t* statistic (see chapter 4) would be appropriate. As an alternative to the numeric criterion, investigators could use the data from which the criterion was derived. Uncertainty estimates from that data set could be used in statistical comparisons.

Key Elements

Several specific survey designs are appropriate for use in a biocriteria program, but designs for a particular environmental assessment should be developed with the aid of a consulting statistician. Such plans should include the following key elements, beginning with the notion of a pilot study.

Pilot Studies

In a pilot study, the investigator makes a limited survey of the variables that determine impact at both the impact and control site. Data from the survey can be used to estimate sample sizes, evaluate sampling methods, establish important variance components, and critique or reevaluate the larger design. The sample size helps determine the particular levels of statistical confidence that can be gleaned from the study. In general, a pilot study can save time and effort by verifying an investigator's preliminary assumptions and initial evaluations of the impact site. Current studies

and historical data collected at the site of interest or similar sites can be used to help establish a good monitoring design.

Location of Sampling Points

A second key issue in the study design is the location of the sampling points. Many specific designs and variations are available, including (1) completely random sample designs, (2) systematic sample designs, and (3) stratified random sample designs.

■ **Random Samples**. In complete random sampling, every potential sampling point has the same probability of selection. The investigator randomly assigns the sample points within the impact site and independently within the control site. No attempt is made to partition the impact and control sites either spatially or temporally except to ensure similar physical habitats. The sampling units are numbered sequentially, and the selection is made using a random number table or computer-generated random numbers.

The advantage of random sampling is that statistical analysis of data from points located completely at random is comparatively straightforward. In addition, the method provides built-in estimates of precision. On the other hand, random sampling can miss important characteristics of the site, spatial coverage tends to be nonuniform, and some points may be of little interest.

■ **Systematic Samples**. Systematic sampling occurs when the investigator locates samples in a nonrandom but consistent manner. For example, samples can be located at the nodes of a grid, at regular intervals along a transect, or at equally spaced intervals along a streambank. The grid or interval can be generated randomly, after which the position of all samples is fixed in space.

Systematic sampling has two advantages over simple random sampling. First, it is easier to draw, since only one random number is required. Second, the sampling points are evenly distributed over the entire area. For this reason, systematic sampling often gives more accurate results than random sampling, particularly for patchy environments or environments with distinct discontinuous populations.

Systematic sampling also has its disadvantages. For example, if the magnitude of the biological variable exhibits a fixed pattern or cycle over space or time, then systematic sampling is unlikely to represent variance of the entire population. Suppose an organism has several hatches, roughly at equally spaced time intervals during the sampling period, then samples taken at fixed-time intervals may provide a biased estimate of the average number of individuals alive at one time. If possible, the population should be checked for such periodicity. If periodicity is found or suspected but not verifiable, systematic sampling should not be used.

Another disadvantage of systematic sampling is that it is more complicated to estimate the standard error than if random sampling had been used. Despite these problems, systematic sampling is often part of a more complex sampling plan in which it is possible to obtain unbiased estimates of the sampling errors.

■ **Stratified Random Samples**. Stratified samples combine the advantages of random and systematic sampling. Stratified random sampling consists of the following three steps: (1) the population is divided into a number of parts, called strata; (2) a random sample is drawn independently in each stratum, and (3) an estimate of the population mean is calculated. Thus:

$$\bar{y}_{st} = \frac{\sum N_h \bar{Y}_h}{N} \qquad (3.1)$$

where $\bar{y}_{st}$ is the estimate of the population mean, N_h is the total number of sampling units in the h^{th} stratum, and $\bar{y}_h$ is the sample mean in the h^{th} stratum, and $N = \sum N_h$ is the size of the population. Note that N_h are not sample sizes but the total sizes of the strata, which must be known to calculate this value.

Stratification is employed if it can be shown that differences between the strata means in the population do not contribute to the sampling error in the estimate of $\bar{y}_h$. In other words, the sampling error of $\bar{y}_h$ arises solely from variations among sampling units that are in the same stratum. If the strata can be formed so that they are internally homogeneous, a gain in precision over simple random sampling can occur.

In stratified sampling, the sample size can vary independently across strata. Therefore, money and human resources can be allocated efficiently across strata. As a general rule, strata with the greatest uncertainty (i.e., with the largest expected variance, or about which little is known) should receive the greatest amount of sampling effort.

For environments that are known to be fairly homogeneous with respect to the biological variable under consideration, stratified random sampling will not add precision to the population estimates. In fact, using stratification in these environments may introduce a loss of precision and a possible bias in the population estimates. In these cases, the investigator may save a great deal of time and effort by using simple random sampling in the sampling plan.

Location of Control Sites

Under EPA's biocriteria program, states may establish either site-specific reference sites or ecologically similar regional reference sites for comparison with impacted sites (U.S. Environ. Prot. Agency, 1990). Typical site-specific reference sites may be established along a gradient. For example, a reference site can be established upstream of a wastewater outfall (Fig. 3.1). Gradients work well for rivers and streams; for larger waterbodies, reference sites can be established on a one-to-one basis with a similar waterbody in the region not experiencing the impact under evaluation.

An important consideration in site-specific reference conditions is to establish that the control site is not impaired at all or that it is only minimally impaired. In particular, baseline data should be obtained to demonstrate that the impact is linked to the differences detected between the reference site and the control site.

Ideally, a reference site should exhibit no impairment; however, natural variability in biological data may make the determination of minimal or no impact difficult, especially if the impact is relatively small. An interesting method for site selection is to establish several reference sites based on their physical similarities with the impact site. For example, selecting one reference site with higher flow than the impact site and another with lower flow may increase the investigator's ability to determine the presence of a real impact. Comparisons of data collected from the impact and reference sites should provide consistent interpretations of the impact, regardless of which reference site is used in the comparison.

Minimizing temporal variation in biological measurements can be critical to the evaluation of control and impacted sites. A general rule is that samples should be obtained from the control and reference sites during the same time periods. It may be feasible to target an index period (e.g., late spring or summer) in which the biological variables are assumed to be appropriate indicators of ecological health (e.g., the period of maximum abundance or the period of minimum variation in water chemistry). However, for some organisms, periods of maximum abundance may also be periods of high variability. In this case, periods of low abundance but stable conditions can be used to help the investigator detect impairment if it exists.

Estimation of Sample Size

A final key component in developing a survey design is to determine how many samples are required. In most plans, the issue involves a trade-off between the accuracy of the sample estimate and the magnitude of available monetary and human resources. Consequently, the first step is to determine how large an error can be tolerated in the sample estimate. This decision requires careful thought; it depends on how the collected data will be used and the consequences of a sizable uncertainty associated with the sample estimates. Thus, in reality, selecting a sample size is somewhat arbitrary and driven by practical considerations of time and money. Investigators should, however, always approach the selection of sample size using sound statistical principles.

The appropriate equations for calculating sample sizes are often design dependent. Here, we present a design for simple random sampling. Suppose that d is the allowable error in the sample mean, and the investigator is willing to take only a 5 percent chance that the error will exceed d. In other words, the investigator wants to be reasonably certain that the error will not exceed d. The equation for the sample size is

$$n = \frac{t^2\sigma^2}{d^2} \qquad (3.2)$$

and t is the t statistic for the level of confidence required. For a 95 percent confidence level that the sample mean will not exceed d, $t = 1.96$. Obviously, an estimate of the population standard deviation, σ, is necessary to use this relationship. In many cases, an estimate of σ can be obtained from existing data. When few data are available about σ, it is a good idea to generate a set of tables to develop a sense of the range of samples required.

Suppose, for example, that an investigator wishes to estimate mean pH readings above a wastewater discharge. How many samples are needed to estimate the true mean pH? At the extremes, the investigator guesses that the standard deviation might range between 0.5 and 1.2 pH units. This estimate leads to Tables 3.1 and 3.2:

Table 3.1.—Number of samples needed to estimate the true mean (low extreme).

CONFIDENCE LEVEL	MARGIN OF ERROR (σ=0.5)		
	0.2 pH units	0.5 pH units	1 pH unit
95%	24	4	1
90%	17	3	1

Table 3.2—Number of samples needed to estimate the true mean (high extreme).

CONFIDENCE LEVEL	MARGIN OF ERROR (σ=1.2)		
	0.2 pH units	0.5 pH units	1 pH unit
95%	138	22	6
90%	98	16	4

Note that the number of required samples increases dramatically as the confidence and precision in the estimates increase, and as the population standard deviation increases. As a general rule, the precision of the estimate is inversely proportional to the square root of the sample size. Therefore, increasing the sample size from 10 to 40 will roughly double the precision.

For a fixed precision, changing the required confidence in the estimate from 95 to 99 percent slightly more than doubles the sample size. Equation 3.2 can easily be adopted for binary response variables in which the responses are expressed as proportions or percentages (Cochran, 1963). In addition, for those situations where the number of sampling units is finite, a finite population correction for the sample size is available (Cochran, 1963).

Equations for calculating sample sizes for random, nonrandom, and stratified sample surveys can be found in the literature. They depend on the sample design, the available variance estimates, and whether the environmental assessment has both spatial and temporal components.

Important Rules

Developing a sample design is frequently driven by factors other than statistics and biology. For example, the investigator may be asked to determine a difference between upstream and downstream stations of a municipal treatment plant outfall, long after the suspected impacts began. Even in these cases, creative sampling strategies can help develop the link between the wastewater outfall and downstream impacts. The following rules apply to most environmental assessment scenarios.

- **Rule 1**. Sample designs and their associated analytical techniques can be difficult to conceptualize and implement. Always consult individuals with appropriate training before starting a biocriteria study.
- **Rule 2**. State precisely and clearly the problem under evaluation before attempting to develop a survey design.
- **Rule 3**. Collect samples from a reference site as a basis for inferring impact. In general, the sampling scheme used at the impacted site should be the same as that employed at the reference site.
- **Rule 4**. To the degree possible, use environmental characteristics to minimize the error in the sample estimate. For example, for patchy environments examine the possibility of systematic sampling; for heterogeneous populations, examine the possibility of using stratified random sampling. In all cases, attempt to minimize sample bias by randomly allocating samples (either geographically or temporally across the entire population, or within strata).
- **Rule 5**. For seasonally dependent biocriteria, collect data for several seasons before attempting to determine an impact. For biocriteria that are not seasonally dependent, collect sufficient data to represent the variability in the population.
- **Rule 6**. Collect enough data so that the accuracy and precision requirements associated with using the information are achieved.